Inference of the Russian Drug Community from One of the Largest Social Networks in the Russian Federation

L.J. Dijkstra · A.V. Yakushev · P.A.C. Duijn · A.V. Boukhanovsky · P.M.A. Sloot

10 December 2012

Abstract
Objectives This study aims to gain insight into what constitutes the drug community in the Russian Federation; information that is absent in official governmental data but is vital for developing effective and much needed intervention strategies to counter the on-going 'drug epidemic'.
Methods Members of the on-line drug community are identified from a crawled set of almost 100,000 users from the social network *'LiveJournal'* by context sensitive text mining of the users' blogs using a dictionary of known drug-related official and 'slang' terminology. The interests that are more (or less) common within this sub-community are determined using Fisher's exact tests and Hochberg and Benjamini's false discovery rate control procedure. A 'psychological portrait' of the 'average' Russian drug user is created by clustering these indicative interests. In addition, a naive Bayesian classifier is presented for assessing one's susceptibility to the 'drug virus'.
Results A total of 268 significant interests separating between users that most actively spread information on narcotics and the rest of the network and a set of themes summarizing these interests. Three sub-networks of users which can be uniquely classified as being either 'infectious', 'susceptible' or 'immune' to the 'drug virus'.
Conclusions The 'average' drug user in the Russian Federation is generally more interested in topics such as Russian rock, non-traditional medicine, UFOs, Buddhism, yoga and the occult. The three sub-networks are all scale-free. The presented method seems to be fruitful for assessing opaque communities within society.

Keywords Illicit drug use · Social network · LiveJournal · Power-law · Russian Federation

1 Introduction

Since the fall of the Soviet Union in the early nineties drug abuse has seen a dramatic increase in the Russian Federation. From 1990 to 2001 the number of registered drug addicts and drug-related crimes went up a nine- and fifteenfold respectively (Sunami, 2007) and continued to rise over the last decade (Mityagin, 2012). The rapid spread and extent of this 'drug epidemic' is of immediate concern to the Russian government and finding effective ways to halt this trend is considered to be of outmost importance.

Due to the criminal nature and general social disapproval of drug use it is complicated to assess the drug community directly. Official governmental statistics do provide an insight into the general trend, but only manage to scratch the surface of the entire drug community in the Russian Federation. The drug users registered in their databases are often among the extreme cases: they have been in one (or more)

L.J. Dijkstra and A.V. Yakushev share first authorship of this work.

L.J. Dijkstra · P.M.A. Sloot
Computational Science Department, University of Amsterdam (UvA), The Netherlands.
E-mail: louisdijkstra@gmail.com

A.V. Yakushev · L.J. Dijkstra · A.V. Boukhanovsky · P.M.A. Sloot
High-Performance Computing Department, National Research University of Information Technologies, Mechanics and Optics (NRU ITMO), Saint Petersburg, Russia.
E-mail: andrew.yakushev@yandex.ru

P.A.C. Duijn
Criminal Intelligence Analysis, Dutch Police, The Netherlands.
E-mail: pacduijn@gmail.com

A.V. Boukhanovsky
E-mail: avb_mail@mail.ru

P.M.A. Sloot (corresponding author)
School of Computer Engineering (SCE), Nanyang Technological University (NTU), Singapore.
Tel.: +31 (0) 20 525 7537
Fax: +31 (0) 20 525 7419
E-mail: p.m.a.sloot@uva.nl

rehabilitation programs or were arrested for using and/or selling illicit narcotics. The (still) 'moderate' user stays out of the picture, making it difficult to obtain reliable information on the drug community as a whole. Within criminological research this non-registered crime is often referred to as *dark number*, see Coleman and Moynihan (1996), and Rhodes et al. (2006).

Gaining a better understanding of what constitutes the drug community in the Russian Federation and in which ways its members can influence (or even inspire) others to start using might prove valuable for devising more effective intervening strategies that can turn the current situation for the better.

In order to handle the drug society's inherent complexity, we will partition the Russian population into (roughly) three groups varying in their involvement in illicit drug use:

1. The *immune*: the group of people that because of, for example, social commitments (e.g., marriage, children, job) and/or strongly held (religious) convictions will not be persuaded to start using drugs.
2. The *infectious*, i.e., the drug community: the group consisting of all individuals involved with drug abuse in one way or another (i.e., using, selling or producing).
3. The *susceptible* containing all individuals that are not a member of one of the previously mentioned groups. They are not involved in any way with illicit drug use at the moment, but might, due to their social position and environment be drawn toward drug use in the future.

The idea to divide the population into these three groups was inspired by the division often used in models for virus spread, see for example the SIR-model of Daley and Kendall (1964), since a similar process seems to underlie the spread of drug addiction through society: infectious (drug users/dealers) can infect susceptible others with the (drug) virus by means of direct and personal contact (i.e., sharing or selling drugs). This analogy has been made before, not only between virus spread and drug addition (Agar, 2005; Beenstock and Rahav, 2004; Mityagin, 2012), but also in the field of 'obesity spreading' (Gallos et al. 2012) and for modeling the spread of information (Iribarren and Moro, 2009; Onnela et al., 2007; Bernardes et al., 2012).

Social network sites (SNSs) have proved over the years that they provide means to uncover social structures and processes that were difficult to observe before (Scott, 2011). In this paper we investigate the social network site *LiveJournal*[1]. With approximately 2.6 million registered Russian users and over 39 million registered users worldwide, it is one of the largest and most popular SNSs in the Russian Federation. The site offers its users an easy-to-use blog-platform where people can read and share their articles with others. In contrast to micro-blogging SNSs such as Facebook[2] (Wilson et al., 2012; Ferri et al., 2012) or Twitter[3] often mentioned in the literature, the site offers a tremendous amount of large user-written texts, making it extremely suitable for text-mining and, consequently, a unique source of data. Maybe because of having the impression to be among 'friends', LiveJournal users write sometimes quite openly about their personal lives in their blogs. Some even comment on their use of drugs and their experiences with various kinds of narcotics. Others (the extreme cases) describe in detail the production process. These openly online expressions can be ascribed to the *on-line disinhibition effect* (Suler, 2004); the invisible and anonymous qualities of online interaction lead to disinhibited, more intensive, self-disclosing and aggressive uses of language. Furthermore, recent studies show that criminal organizations are actively using on-line communities as a new 'business' tool for communication, research, logistics, marketing, recruitment, distribution of drugs and monetarization (Décary-Hétu and Morselli, 2011; EUROPOL, 2011; Walsh, 2011; Choo and Smith, 2008; Williams, 2001). Research of on-line communities, therefore, might aid in gaining a better understanding of the behavior of opaque networks within a society.

In order to get a better insight into the drug community in the Russian Federation, we crawl a large randomly selected group of Russian LiveJournal users. Every blog entry of every user is associated with a weight indicating to what extent it refers to illicit forms of drug use by overlaying the document word-for-word with a dictionary consisting of known drug-related terminology (both official as well as informal/'slang'). When the sum of 'indicator' weights of all the blog entries of a specific user reaches a certain threshold, the user is considered to be a member of the on-line drug community. The idea behind this approach is that drug users are more likely to use drug-related terminology in their blog entries than others. We will return to this assumption extensively in Section 5. The way users are classified and the drug-dictionary are discussed in detail in Section 3.2.

After identifying the on-line drug community, we might ask ourselves what kind of people are generally to be found in this sub-network? In order to get a better picture of the 'average' user in this sub-community, we gather all the interests mentioned on each user's profile page and compare how often they appear within the on-line drug community with the frequency of appearance in the rest of the network. We limit ourselves here to interests, due to the fact that it is rather unclear how to automatically construct a 'psychological profile' of a user based solely on his or her texts. That way, we try to isolate those interests that are truly more common in one of these two distinct groups of users. In Section 3.3 we describe the used methodology in more detail.

[1] LiveJournal is available at `http://www.livejournal.com` (English) and `http://www.livejournal.ru` (Russian).

[2] Facebook is available at `http://www.facebook.com`.

[3] Twitter is available at `http://www.twitter.com`.

The susceptibility of people to the 'drug virus' is thought to depend on their exposure to drug-related information and their own interest in this topic. This social mechanism of transmission is called *differential association* in which drives, techniques, motives, rationalizations and attitudes toward deviant behavior are learned and exchanged by social interaction (Sutherland, 1947; Lanier and Henry, 1998; Haynie, 2002). From this perspective the number of interests a user has in common with the on-line drug community might indicate a higher susceptibility, since 1) this person is more likely to stumble upon blog entries published by member of the on-line drug community (which are more often about drug use), and 2) it might indicate a certain lifestyle more prone to drugs. Following this reasoning, we present a naive Bayesian classifier using the log-likelihood ratio method (Kantardzic, 2011; Hastie et al., 2009) in Section 3.4 that assesses the susceptibility of a user to drugs given his/her personal interests. When a user's interests overlap more with the interests in the on-line drug community than the interests of the rest of the population, they are considered to be susceptible.

Users that were not identified as being a member of the on-line drug community on the basis of their written texts or as susceptible due to a large similarity with their interests and the interests common in the on-line drug community are considered to be immune. They do not write (much) about illicit drug use and their interests do not suggest a lean towards the on-line drug community.

After having (roughly) identified the three subgroups (i.e., immune, infectious and susceptible) in the social network LiveJournal, we might wonder whether there are structural differences between the corresponding subnetworks. In Section 4.3 we will describe and compare them.

The remainder of this paper is organized as follows. In Section 2 we discuss the social network site LiveJournal, describe the kind of information users put out about themselves and point to several unique features this SNS has over others often studied in the literature. Section 3 describes the crawled LiveJournal data set and the methods used to partition its users and determine significant interests. The results are presented in Section 4. We will finish with several conclusions, a rather extensive discussion and a few pointers for future research. In Appendix 1 we explore the frequency with which interests appear in the network and show that this probability distribution follows a power-law.

2 The SNS LiveJournal

The social network site LiveJournal with over 39 million worldwide and approximately 2.6 million registered Russian users is by far the most popular blog-platform in the Russian Federation. With 1.7 million active users and (approximately) 130,000 new posts every day the site offers a fast body of data for studying social structures and processes[4]. In addition to publishing their own articles, the users are offered the possibility to enter information on their whereabouts (e.g. hometown), demographics (e.g. birthday), their personal interests (e.g. favourite books, films and music) and even their current mood (e.g. happy, sad). Articles can be tagged and an extensive comment system provides the readers with the possibility to respond and exchange opinions and ideas.

Users can unilaterally declare any other registered user as a 'friend', i.e., ties are unidirectional. A tie reflects the desire of a user to keep up-to-date with the articles of the other. Consequently, every profile contains two lists of ties: 1) a list of alters that currently follow the articles published by the ego, and 2) a list of alters whose articles the ego follows. (Note the similarity with Twitter). We will refer to these lists as the list of *followers* and *following friends*, respectively.

LiveJournal differs from other (large) social network sites in two important aspects: 1) it has a large number of users that actively write in Russian, and 2) the texts are large in contrast to the micro-blogging SNSs often considered in the literature (Wilson et al., 2012). The latter makes LiveJournal exceptionally suitable for text-mining and, as such might provide insights into social structures and processes where other SNSs cannot.

3 Methods

Section 3.1 describes the data collected from the SNS LiveJournal. In Section 3.2 we discuss the drug-dictionary and procedure used for classifying those users who are most likely to be involved in drug abuse. After colouring the subnetwork of the on-line drug community, we proceed in Section 3.3 with identifying those interests that are more common for this set of users or the rest of the on-line. These indicative interests are used by the nave Bayesian classifier introduced in Section 3.4 for identifying the 'susceptible' and 'immune' subnetworks. We will later analyze the structure of these three subnetworks later in Section 4.3.

3.1 The LiveJournal Data Set

On the 9th of September 2012 we crawled 98602 randomly selected Russian user profiles. For each profile we stored its username, the last 25 posted blog entries, personal interests and the lists of followers and following 'friends'. In addition we stored (when available) the user's birthday and place of living.

In order to collect this data, we developed a distributed crawler that employs the MapReduce Model (Lämmel, 2007)

[4] LiveJournal's own statistics page can be found at `http://www.livejournal.com/stats.bml`.

and the open source framework Apache Hadoop (White, 2009). The system is similar to the Apache Nutch crawler (Cafarella and Cutting, 2004) but allows for multiple users to collect and process data at the same time; the fetcher module is moved outside the Hadoop framework making it a separate application that can run on various machine architectures simultaneously.

A total of 22357 users fully specified their birthday on their public profile (ages higher than 80 were regarded to be reported falsely). In Section 4.1 we explore some characteristics of the crawled population and compare it with the Russian population.

3.2 The On-line Drug Community

Users are classified as being a member of the on-line drug community by comparing their last 25 blog entries with a dictionary of known drug-related terminology collected by drug experts at the Saint Petersburg Information and Analytical Center[5]. The total of 368 words in this dictionary are split up into two categories: official and informal/'slang' terminology. Official terminology are words that are unmistakingly related to illicit drug use (e.g., cocaine and heroin) and are assigned a high weight, i.e., 5. Informal/'slang' expressions can often be interpreted in various ways and cannot be directly related to drug use. For example, the Russian word 'kolesa' refers normally to wheels while it also can be used (in rather dubious circumstances) as a word for pills. To account for this ambiguity, 'slang' expressions are assigned a lower weight than official terminology, i.e., 1. Table 1 shows a few example words from the dictionary alongside their weight and (free) English translation[6].

In addition to this set of words, each blog entry was also checked for the presence of a collection of drug-related phrases. The presence of certain combinations of words in a text, e.g., 'injecting' and 'heroin', is a strong indication that the author is involved with illicit drug use. In order to account for this valuable information, the dictionary consists additionally of 8359 phrases, each assigned with a slightly higher weight than the mere sum of the words it consists of[7].

In order to compare inflected or derived words in the posts with words in the dictionary we first reduce them to their root form using a Russian version of the Porter stemming algorithm (Porter, 1980; Porter, 2006).

When the summed weights of all the blog entries of a user reaches a certain threshold, he/she is considered to be a member of the on-line drug community. Users who use a small number of the words and phrases from the dictionary in a limited number of blog entries are, thus, less likely to be identified as a member than the ones who frequently use drug-related terminology throughout a large numbers of texts. The threshold was set manually, see Fig. 1.

We will refer to the entire set of users who's summed weights reaches the threshold as the on-line drug community throughout the rest of this paper. To what extent the sub-community corresponds to the Russian drug community will be a point of discussion in Section 5.

Table 1 Examples of words in the drug-dictionary

Russian	English translation	Weight
Kokain	Cocaine	5
Geroin	Heroin	5
Mariguana	Marijuana	5
Abstyag	Withdrawal syndrome	5
Tabletki	Pills	1
Kolesa	Pills/Wheels	1

3.3 Identifying Common Interests of the On-line Drug Community

In this section we will formulate an approach for determining which interests are most common (or uncommon) for a particular subset of SNS users, in our particular case, the on-line drug community.

First, we collect the interests on the profile pages of all users in the on-line drug community that at least appear more than 10 times. (The reason for disregarding rather unfrequent interests is that they do not add much when one wants to gain a better understanding of an entire community). Lets denote this set of interests with $\mathbf{I} = \{I_1, I_2, \ldots, I_m\}$. Since the members of the on-line drug community are known, we are able to count how often users express their interest in both this sub-community and the rest of the social network. For every interest I_i we can, thus, obtain a 2×2 contingency table similar to Table 2 where $(a+b+c+d) = n$ is the total number of users in the crawled population that have at least

Table 2 The 2×2 contingency table for interest I_i

	Drug community	Rest	*Total*
Is interested in I_i	a	b	$a+b$
Not interested in I_i	c	d	$c+d$
Total	$a+c$	$b+d$	n

[5] The homepage of SPb IAC can be found at `http://iac.spb.ru` (in Russian).

[6] The full drug-dictionary is freely available and can be downloaded at `http://escience.ifmo.ru/?ws=sub48`.

[7] The number of phrases (8359) is rather high in comparison to the number of words (368) in this dictionary. This is due to the fact that we consider a phrase consisting, for example, of the words 'injecting', 'heroin' and the phrase with the words 'injection', 'heroin' and 'needle' as two separate expressions (where the latter is associated with a higher weight than the former).

one interest on their profile page (i.e., $n = 62370$), $a+c$ is the number of users identified as members of the on-line drug community, $a+b$ is the total number of users who expressed their interest in I_i and $c+d$ are the users not interested in I_i. The question is whether this interest appears significantly more (or less) in the on-line drug community than in the rest of the rest of the network, i.e., do the proportions $a/(a+c)$ and $b/(b+d)$ differ?

We, thus, have m null hypotheses (H_i^0), one for each interest I_i in $\mathbf{I}$. Applying the two-sided version of Fisher's exact test[8] (Fisher, 1922; Agresti, 1992) to each contingency table provides us with their corresponding p-values: $p_1, p_2, \ldots, p_m$.

The total number of null hypotheses is large (3282 to be precise, corresponding to the total number of interests expressed more than 10 times in the on-line drug community). Simply comparing the obtained p-values with a common fixed significance level (e.g., $p \leq .05$) will result in a high number of false discoveries, i.e., falsely rejected null hypotheses. Benjamini and Hochberg (1995) showed that the expected false discovery rate can be upper bounded by $q \in [0,1]$ with the following control procedure[9] (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001):

1. Order the p-values in increasing order, i.e., $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
2. For a given q, find the largest k for which $p_{(k)} \leq kq$.
3. Reject all $H_{(i)}^0$ for $i = 1, 2, \ldots, k$.

We will use a q-value of 5%. The interests associated with all rejected $H_{(i)}^0$, $\mathbf{I}' = \{I_{(1)}, I_{(2)}, \ldots, I_{(k)}\}$, are considered to be the interests that really differ between the on-line drug community and the rest of the social network.

Due to the large sample size and the initially large number of interests, the number of significant interests in $\mathbf{I}'$ is expected to be quite high. Partitioning them into a set of *themes* might help with getting a better overview of the wide variety of significant interests. In order to do so, we cluster the set of significant interests $\mathbf{I}'$ using a hierarchical agglomerative clustering algorithm with a complete linkage strategy (Kantardzic, 2011; Everitt, 2001). Complete-linkage is preferred here over single-linkage due to the fact is does not suffer from the chaining phenomena, i.e., clusters may be forced together due to single elements being close to each other, even if a majority of elements is very distant. Average-linkage was no option due to its high computational load. The similarity between two clusters of interests, C_1 and C_2, is defined as

$$sim(C_1, C_2) = \frac{n(S_1 \cap S_2)}{\sqrt{n(S_1) \cdot n(S_2)}} \quad (1)$$

where S_1 and S_2 are the sets of users that expressed their interest in at least one of the topics in, respectively, C_1 and C_2. $n(\cdot)$ returns the number of users. This similarity measure is known as *cosine similarity* or more commonly known in biology as the Ochiai coefficient (Ochiai, 1957). We will refer to the resulting clusters of significant interests as themes throughout the rest of this paper.

3.4 Assessing Susceptibility

A large number of common interests between a user and the on-line drug community might indicate a higher susceptibility to drugs, since 1) the user is more likely to stumble upon blog entries published by members of this sub-community, and 2) it might indicate a certain lifestyle more prone to drug use. Certain interests might, on the other hand, indicate a low susceptibility. Think of interests that suggest that the user in question has certain social commitments (e.g., marriage, children, job) or strong-held (religious) convictions. The idea that interests are related to susceptibility underlies the classification method in this section: an individual is considered to be a susceptible user when his/her personal interests resemble the interests common for the drug community more than the interests of the rest of the on-line population.

A naive Bayesian classifier was used (Kantardzic, 2011). Due to the fact that certain combinations of interests are rare, we are forced to assume conditional independence between each pair of interests and use the log-likelihood ratio method.

Let us first define k feature variables, one for each interest in the set $\mathbf{I}'$:

$$\mathbf{F} = \{F_1, F_2, ..., F_k\}$$

where F_i is true when the user is interested in I_i in $\mathbf{I}'$ and otherwise false. The set of feature variables $\mathbf{F}$ is used to describe the personal interests of each user in the network.

The chance that a user belongs to the drug community (D) given his/her interests is given by the conditional chance $P(D \mid \mathbf{F})$. Given the assumption that each feature variable F_i is conditionally independent of F_j when $i \neq j$, i.e., $P(F_i \mid D, F_j) = P(F_i \mid D)$, this probability can be expressed as

$$P(D \mid \mathbf{F}) = \frac{P(D)}{P(\mathbf{F})} \prod_{i=1}^{k} P(F_i \mid D). \quad (2)$$

Similarly, the chance of not being a member of the drug community given the users interests is

$$P(\neg D \mid \mathbf{F}) = \frac{P(\neg D)}{P(\mathbf{F})} \prod_{i=1}^{k} P(F_i \mid \neg D). \quad (3)$$

[8] A χ^2 test originally designed for 2×2 contingency tables by Sir R.A. Fisher (1922).

[9] Strictly speaking, the expected false discovery rate is only upper bounded when the m test statistics are independent, which does not hold in this particular case. B. Efron makes the case in his book *Large-Scale Inference* (2010) that this independency constraint is not strong.

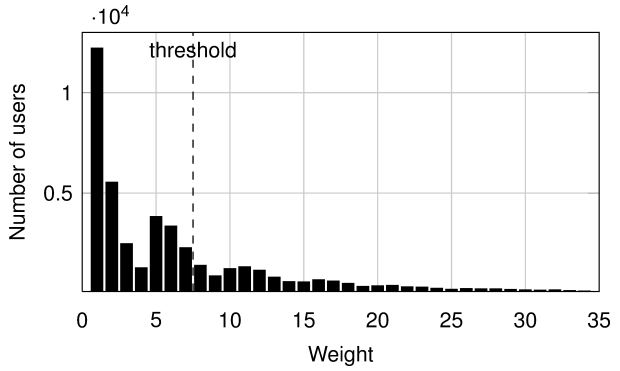


Fig. 1 The summed weights of the blog entries of each user in the LiveJournal data set. The higher the summed weight the more the user used the words and phrases present in the drug-dictionary (see Section 3.2). Users are considered to be a member of the on-line drug community when their weighted sum crosses the threshold of 8

By applying the log-likelihood ratio method, i.e., dividing eq. (2) by eq. (3) and taking the natural logarithm of both sides, we find that the inequality $P(D \mid \mathbf{F}) > P(\neg D \mid \mathbf{F})$, i.e., the user is more likely to belong to the drug community given the user's interests, is equivalent to the inequality:

$$\log \frac{P(D)}{P(\neg D)} + \sum_{i=1}^{k} \log \frac{P(F_i \mid D)}{P(F_i \mid \neg D)} > 0. \tag{4}$$

A user is considered to be susceptible when he/she does not belong to the drug community and this inequality holds. Users that are not a member of the on-line drug community or considered to be susceptible, are immune.

4 Results

In order to identify those users in the network involved with illicit drug use, we overlaid their last 25 blog entries with a dictionary of known drug-related terminology (see Section 3.2). Fig. 1 depicts the distribution of the weights assigned to the randomly crawled LiveJournal users. Note that the majority of users appear to make use of a rather small number of drug-related terminology. The fluctuations that can be seen around the weights 5, 10 and (less distinct) 15 and 20 can be explained by the weights assigned to the words present in the drug-dictionary (5 for official, clearly drug-related, terminology and 1 for (ambiguous) 'slang' expressions). The users with the highest weights are assumed to be the ones most interested and/or involved in illicit drug use. The threshold was set to 8 (see Fig. 1), i.e., when the weight of a user crosses 8, he/she is considered to be a member of the on-line drug community. Other thresholds close to 8 were considered as well. We found that the themes as presented in Section 4.2 did not change tremendously. By setting the threshold to 8, approximately 20% of the total set of crawled users were classified as being a member of the on-line drug community.

4.1 Characteristics of the SNS LiveJournal

Fig. 2a depicts the age distribution of the LiveJournal data set split out between the on-line drug community and the susceptible and immune user groups. Note that this SNS is especially popular among 20 to 40 year old individuals. Figure 2b depicts the age distribution of the Russian Federation as determined on the 1st of January 2011. The data was made available by Rosstat[10]. The major dip around the ages 62-70 is a reflection of the impact that the Second World War had on the Russian population.

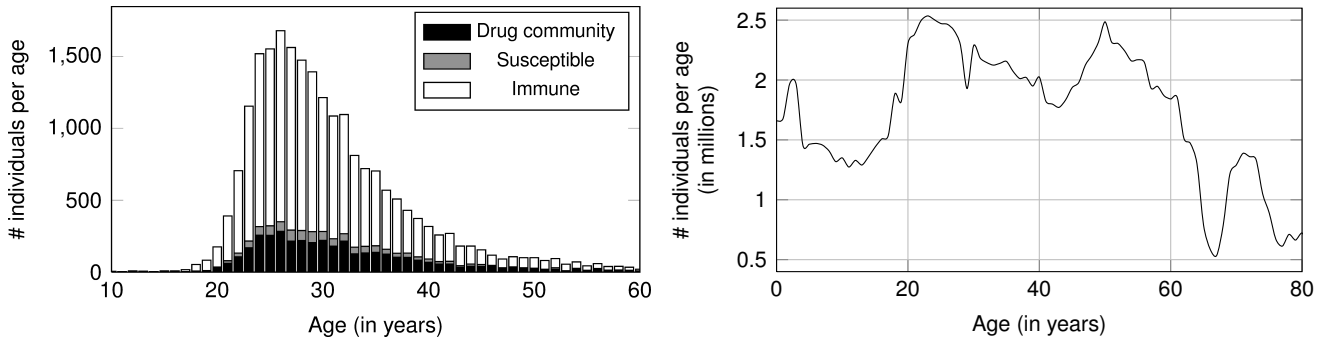
Note the difference between the Russian LiveJournal community and the Russian population as a whole. Using LiveJournal to sample the Russian population poses two problems: 1) one only samples those individuals who are registered as a user in this SNS, and 2) we seriously oversample the age group 20-40. Both aspects might not pose a real threat; the Russian drug community is, as mentioned before, difficult (or even impossible) to sample directly, making sampling a SNS one of the limited options one has, when one wants to gain a better insight into this sub-community. In addition, illicit drug use is known to occur especially in this particular age group (Mityagin, 2012). The strong presence of this group, thus, might help in gathering more information on the community of interest.

Of the total number of 98602 users studied in the LiveJournal data set, 16553 and 3586 were identified as, respectively, members of the drug community and susceptible users. Susceptible users are identified using the naive Bayesian classifier as described in Section 3.4 which makes use of the interests the user posted on his/her profile page. Common interests can be shown to be rare. In fact, the frequency with which an interest is mentioned by users of this SNS can be shown to follow a power-law distribution with coefficient $\gamma \approx 1.54$, see Appendix 1. With a low number of common interests, there is often not enough to go on in order to reliably classify a user as being susceptible, which explains the relatively small number of susceptible users found.

4.2 Drug Indicators

After applying Fisher's exact test and Benjamini and Hochberg's false discovery rate control procedure with a q-value of 5% (see Section 3.3), we found 268 of the 3282 initial interests to be significant, i.e., the on-line drug community is, thus, more/less interested in these topics than the rest of the LiveJournal users. In order to assess to what extent an interest I is indicative for being a member of the drug community (D) or the rest of the population, we use the conditional probability $P(D \mid I)$. Among the interests most indicative for the

[10] The governmental statistics agency of the Russian Federation. They can be found at `http://www.gks.ru` (in Russian) with links to their rather extensive database.



a LiveJournal (2012) **b** Russian Federation (2011)

Fig. 2 a The age distribution of the LiveJournal data set (2012) split out between the on-line drug community, and the susceptible and immune user groups. Note that this SNS is especially popular in Russia among 20 to 40 year olds **b** The age distribution of the Russian Federation on the 1st of January 2011 (the data was made available by Rosstat). Note the difference between the two age distributions. LiveJournal does, thus, not provide a good sample of the Russian population, although, while investigating illicit drug use it might be useful to sample especially that fraction of the population known to be most involved with narcotics (Mityagin 2012)

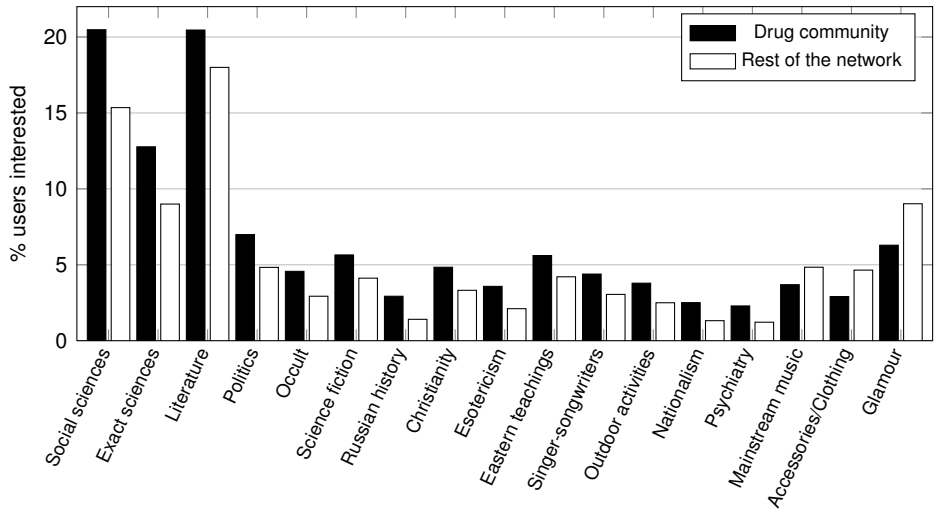


Fig. 3 The percentage of users within the on-line drug community and the rest of the on-line population interested in each theme (see Table 3). Users are considered to be interested in a theme when they mention at least one of the interests contained in that theme. Note that the last three themes are more likely to be found in the non-drug section of the network. The other themes are relatively more likely to appear in the on-line drug community

on-line drug community (i.e., $P(D \mid I) > .5$), we found interests such as: the White Movement (a loose confederation of anti-communist forces who fought the Bolsheviks in the Russian civil war; now often associated with the Russian nationalistic movement), humanistic psychology, partisans, Aryan (ancient people that partly inhabited current Russian territory), Stalinism, Dadaism, narcology and Magadan (a city in the far east of the Russian territory, famous for its large jail). Among interests most indicative for not belonging to the drug community (i.e., $P(D \mid I) < .5$), we found interests such as: accessories, beads, jewellery, London, clothing, glamour, handmade, shoes, beach and interior design.

In order to get a better view on the wide variety of significant indicative interests, we clustered them using the cluster algorithm described in Section 3.3. We found 42 different themes in total. In this Section we will only discuss the ones most prominent within the on-line drug community and the rest of the LiveJournal population.

Fig. 3 shows the various themes and to what extent they appear in the on-line drug community and the rest of the LiveJournal population. We consider users to be interested in a theme, when they mention at least one of the interests contained in that theme on their profile page.

The names assigned to each theme were determined by the writers of this article. In order to overcome some of the inevitable subjectivity inherent to this process, we will describe the themes shortly in Table 3, where the second column denotes the number of significant interests in each theme. When the number of interests in a theme is small, we

will sum up all the interests (translated to English); otherwise we will suffice with a short description.

In both Fig. 3 and Table 3 the last three themes (Mainstream music, Accessories/Clothing and Glamour) appear more often in the non-drug section of the network. The others are more common for the on-line drug community.

Recall that significant interests were clustered solely on the basis of their cosine similarity (i.e., the more users that expressed their interest in both topics, the higher the 'similarity'). In which of the two distinct communities the interest is more prominent is not taken into account. Each theme is, thus, likely to contain interests that are more common for the on-line drug community and interests that are more often found in the rest of the network. To what extent a theme can be related to one of these two groups can, therefore, be expected to be less clear than for individual interests.

4.3 Network Structure Analysis

In this section we will explore the structure of the on-line drug community, susceptible and immune subnetworks.

Fig. 4a shows the degree distribution of the total crawled LiveJournal network. Degree is defined here as the number of followers and following 'friends' of a user. Note that the number of users seems to decrease exponentially with degree; an indication that the distribution might follow a power-law:

$$p(x) = Cx^{-\gamma} \tag{5}$$

where x is the degree of a user, γ is the power-law coefficient and C is a constant. Power-law distributions appear in a wide variety of natural and man-made processes, e.g., the number of inhabitants in cities, the diameter of moon crates and the intensity of solar flares. The wide-spread appearance of the power-law raises the question whether the same process might underlie these (at first glance) different phenomena, causing quite a discussion in the literature. For a more elaborate discussion of power-laws and their appearance, we refer the reader to a recent paper by Pinto et al. (2012).

Fig. 4b shows the rank/frequency log-log plot[11] of the degree distribution in 4a. Note the points in this plot lie (approximately) on a straight line, which is a characteristic of power-law distributions.

Very few real-word networks display a power-law distribution over the entire degree range, making it necessary to determine where the degree distribution is most likely to start following a power-law (denoted here with x_{min}). The power-law exponent γ and x_{min} were determined using the maximum likelihood method as described in the paper by Clauset et al. (2009) and were found to be equal to 1.54 and 8, respectively. The fit is shown in Fig. 4b as a dashed line. Note that the line seems to fit the data quite well. The standard statistical test for the quality of fit as proposed by A. Clauset, C. Shalizi and M. Newman (2009) shows that the data gives no raise to believe that the degree distribution does not follow a power-law (i.e., $p = .57$ with 1000 repetitions).

Fig. 5 shows the rank/frequency log-log plots of the degree distributions of the on-line drug community, susceptible and immune network together with their power-law fits. Note that these sub-networks also follows a power-law distribution, only with slightly different γ's.

Table 4 shows various characteristics of the LiveJournal network and its three subnetworks. Standard deviations are reported between parentheses. Note that the mean age does not differ much. The large differences between the maximum degrees of these networks are common for heavy right-tailed distributions. The best fits for γ, x_{min} and the p-value of the goodness of fit test are reported as well.

5 Conclusions/Discussion

Drug abuse has seen a dramatic increase in the Russian Federation during the last two decades (Sunami, 2007; Mityagin, 2012). The rapid spread and extent of this 'drug epidemic' forms a serious cause for alarm and finding effective ways to halt the current trend is of outmost importance.

Due to the criminal nature and the general social disapproval of narcotics, it is difficult (or outright impossible) to assess the drug community directly. Official governmental statistics do provide some insight, but fail to give the complete picture; the 'moderate' drug user is hardly noticed. Information retrieved from social networks such as LiveJournal can, therefore, contribute in gaining a better understanding of what constitutes the drug community in the Russian Federation and might prove to be vital for devising more effective intervention strategies.

In this paper we present a method to assess this non-directly observable community by mining the popular social network site LiveJournal. By comparing the users' blogs with a dictionary consisting of known drug-related Russian terminology, we were able to identify those users that write most actively about drug use. By collecting their interests, we were able to create a general picture of the kind of users that can be found within the on-line drug community, see Table 3 and Fig. 3. In addition, we introduced a naive Bayesian classifier for identifying potentially susceptible users by comparing their personal interests with the interests most common within the on-line drug community. The 'infectious', 'susceptible' and 'immune' subnetworks were shown to have a similar structure; their degree distributions follow a power-law, although with slightly varying exponents.

[11] A rank/frequency log-log plot is the plot of the occurrence frequency versus the rank on logarithmically scaled axes. For a more elaborate description on how to construct such a plot, see the paper by Mark Newman (2005), Appendix A.

Table 3 Description of the most prominent themes

Theme	#	Description
Social sciences	5	Sociology, history, economics, psychology and law.
Exact sciences	7	Programming, biology, astronomy, medicine, archeology, ecology and philosophy.
Literature	9	Containing rather general interests such as books, journalism, poetry and prose.
Politics	22	This theme contains various national (opposition, corruption and Russia), international (Chechnya, NATO, Poland and Ukraine) and general (socialism, democracy and anti-communism) political topics.
Occult	15	Concerns a wide variety of topics, including, for example, the occult, non-traditional medicine, mysticism, clairvoyance, telepathy and the prediction of the future through the reading of cards (tarot).
Science fiction	8	Containing interests like UFOs, futurology, nanotechnology, science fiction and the American science fiction writer H. Harrison.
Russian history	11	Ranging from general sciences (anthropology, ethnography, war history) to particular events in the history of Russia (WWII, the Russian civil war) and important historical groups (partizans).
Christianity	3	God, the Russian orthodox church and religion.
Esotericism	7	Contains various topics related to esotericism (esotericism itself, but also the expansion and altering of the human mind) and Castaneda, a rather famous author who popularized topics such as 'stalking' (technique to control the mind) and lucid dreams.
Eastern teachings	10	Various eastern teachings/religions (Buddhism, Zen and yoga) and related terms (e.g., mantras, chakras and tantras).
Singer-songwriters	6	Interests related to Russian rock and singer-songwriters (e.g., V. Vysotsky).
Outdoor activities	8	Diving, fishing, hunting and topics related to Mountain climbing (e.g., alpinism and the Altai mountains) and survival.
Nationalism	9	Covering interests such as the Russian empire, patriotism, the Russian people, the White Movement and antiglobalization.
Psychiatry	6	Including psychiatry, psychoanalysis, psychotherapy, psychosomatic medicine and transpersonal and humanitarian psychology.
Mainstream music	6	Containing several famous mainstream musicians, such as Madonna, Coldplay and Björk.
Accessories/Clothing	13	Varying from accessories like beads, jewelry, shoes and bags to clothing and interior design.
Glamour	13	Includes the interest glamour itself. It further covers fashion (e.g., journals, style, jeans, design and shopping) and the night-life of Moscow.

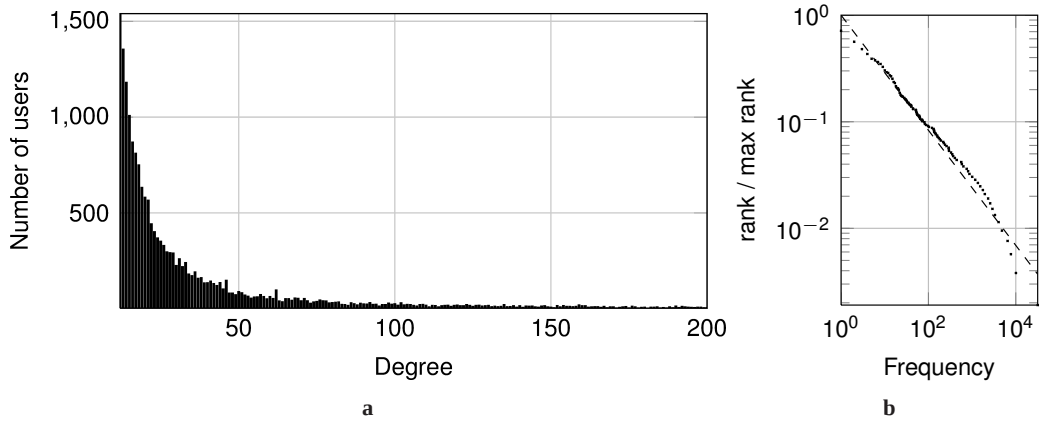


Fig. 4 a A fraction of the degree distribution of the crawled LiveJournal network. Note that the number of users decreases exponentially with degree **b** The rank/frequency log-log plot of the degree distribution and the power-law fit depicted as a dashed line ($\gamma \approx 1.54$ and $x_{min} = 8$). The p-value was found to be approximately .57, i.e., there is no reason to believe that the degree distribution does not follow a power-law

It is unclear to what extent we were able to identify the users that are really involved in drug use. Users that tend to write often about narcotics might do so for the following three reasons: 1) to raise the discussion on the social problems caused by drug abuse or propose possible ways to change the current situation for the better, 2) in an attempt to persuade others to stop or never start using drugs, i.e., 'anti-propaganda', or 3) to share their experiences with drugs or to express their interest in this topic. We are solely interested in the group of users writing about narcotics for the third reason; they are the ones that use drugs or are likely to do so in the future.

The appearance of the theme politics in Fig. 3 might be best explained by the presence of users in LiveJournal that

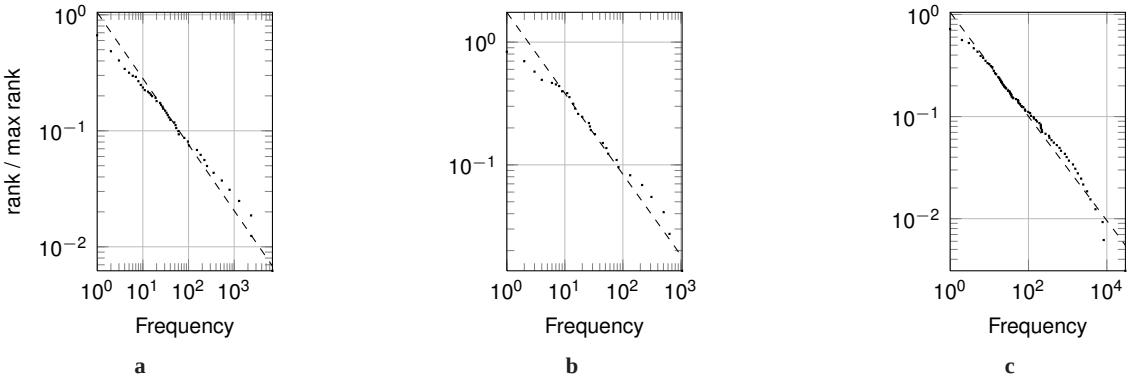


Fig. 5 The rank/frequency log-log plots of the degree distributions of the three subnetworks in the crawled LiveJournal network: the drug community ($\gamma \approx 1.57$ and $x_{min} = 19$) and the susceptible ($\gamma \approx 1.66$ and $x_{min} = 8$) and immune subnetwork ($\gamma \approx 1.54$ and $x_{min} = 10$) . The power-law fit is depicted as a dashed line. The found *p*-values give no reason to believe these distributions do not follow a power-law

Table 4 Structural characteristics of the various subnetworks in LiveJournal

Network	Size	Edges	Age	Max. degree	γ	x_{min}	*p*-value
Drug community	16553	61021	32.08 (9.20)	160	1.57	19	.97
Susceptible	3586	16499	32.14 (8.75)	72	1.66	8	.84
Immune	78463	496018	30.31 (8.03)	323	1.51	10	.76
Total	98602	982197	30.71 (8.32)	524	1.54	8	.57

do not write about drugs because they are personally interested or using them, but rather since they want to bring the social problems related to narcotics under the attention. The same might hold for the themes as the social and exact sciences, psychiatry and, potentially, nationalism. The presence of a theme like Christianity (consisting of the interests 'God', 'the Russian orthodox church' and 'religion') is more likely to be explained by the presence of users that spread anti-propaganda, especially when taking the negative stance of the church towards drugs into account.

Themes such as the occult, esotericism, science fiction and eastern teachings, however, are hardly explained by stating that the users interested in these topics are heavily concerned with the social impact of drug abuse, or actively spreading anti-propaganda. Most likely, we caught a glimpse of the actual drug community.

The explanations of why certain themes are presented in the on-line drug community are, of course, based solely on the view of the authors and, therefore, subjective. Further research is required to establish what themes are truly related to the Russian drug community. In order to establish the validity of the approach described in this paper, one might compare the presented results with law enforcement data, e.g., it would be interesting to compare the number of convictions for drug-related crimes between the on-line drug community and the rest of the crawled LiveJournal population.

The susceptibility of an individual to drugs was determined on the basis of the similarity between his/her personal interests and the interests common in the on-line drug community. We limited ourselves here to their interests, since it was unclear how to relate the susceptibility of a user and his/her texts.

The number of susceptible users is relatively small due to the small number of common interests present in the LiveJournal network. In fact, it can be shown that the frequency with which a certain interest occurs follows a power-law with exponent $\gamma \approx 1.54$, see Appendix 1. With a low number of common interests, there is often not enough to go on to identify a user as being susceptible. It is, thus, very well possible that we overlooked several immune users who should have been noted as being susceptible.

Users were considered to be a member of the on-line drug community when the weighted sum of their blog entries crossed the threshold of 8, see Fig. 1. We experimented with different thresholds and found that, although the list of significant interests does vary, the resulting clusters/themes remain stable. The weights assigned to the official and informal/'slang' terminology in the drug-dictionary were not varied. Since the final themes did not vary much while varying the threshold, it is unlikely that they would now.

As mentioned before, we found that the LiveJournal network and the infectious, susceptible and immune subnetworks are most likely scale-free (i.e., their degree distributions follow a power-law). Although the performed goodness of fit test (Clauset et al., 2009) does not exclude other possibilities, e.g., Poisson, we can state with certainty that the distributions are heavy-right tailed, which entails that

the network has hubs, i.e., users with a far higher degree than the rest of the network. This knowledge might be of major importance when one wants to disrupt the network to, for example, limit the spread of drug-related information on the network. Removing the hubs would heavily disrupt the information flow (Bollobas and Riordan, 2004; Albert et al., 2000; Crucitti et al., 2003).

This paper has shown the promise of 'crawling social networks' in delineating and analyzing social groups that hitherto have eluded such research, because of the fundamentally opaque nature of membership of such groups. The case in point is the Russian drugs community. We hope that continuing research along the lines we set out in this paper will help to map the dynamics of this group, and will ultimately contribute to halting, if not reverting its tragic trend to grow.

Acknowledgements The authors thank Dr. Sergey Mityagin from the Saint Petersburg Information and Analytical Center (SPb IAC) for fruitful discussions on the drug addiction profiles in the Russian Federation. In addition, the authors would like to express their gratitude to Prof. dr. T.K. Dijkstra from the University of Groningen (RUG) and the Free University Amsterdam (VU) for introducing us with false discovery rate control and his useful remarks. This work is supported by the *Leading Scientist Program* of the Russian Federation, contract 11.G34.31.0019, as well as by the Complexity program of NTU, Singapore.

Appendix 1: LiveJournal User Interests

In this appendix we take a closer look at the frequency with which interests are expressed by the users of the social network LiveJournal. Fig. 6 shows the frequency of occurrence of interests within the crawled population. Note that the distribution is heavy right-tailed; its slope suggests that the distribution might follow a power-law, see eq. (5). Fig. 6b shows the corresponding rank/frequency log-log plot of the histogram in 6a. The exponent $\gamma \approx 1.54$ and the start of the distribution $x_{min} = 3$ were approximated using the maximum likelihood method as proposed by Clauset et al. (2009). Note that the fitted line in 6b approximates the distribution quite well. The standard goodness-of-fit test (Clauset et al. 2009) indicates there is no reason to believe that the distribution does not follow a power-law, i.e., the p-value was approximately equal to .57.

The fact that the distribution of interests within the SNS LiveJournal is heavy-right tailed explains why the number of susceptible users (see Table 4) is relatively small compared to the other groups.

References

1. Agar, M. (2005). Agents in living color: towards emic agent-based models. Journal of Artificial Societies and Social Simulation 8(4). `http://jasss.soc.surrey.ac.uk/8/1/4.html`. Accessed 19 November 2012.
2. Agresti A (1992) A survey of exact inference for contingency tables. Statistical Science 7(1):131–177
3. Albert R, Jeong H, Barabasi, AL (2000) Error and attack tolerance of complex networks. Nature 406:378–382
4. Beenstock M, Rahav G (2004) Immunity and susceptibility in illicit drug initiation in Israel. Journal of Quantitative Criminology 20(2):117–142
5. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B (Methodological) 57(1):289–300
6. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Annals of Statistics 29(4):1165–1188
7. Bernades DF, Latapy M, Tarissan F (2012) Relevance of SIR model for real-world spreading phenomena: experiments on a large-scale p2p system. Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM), Istanbul (Turkey)
8. Bollobas B, Riordan O (2004) Robustness and vulnerability of scale-free random graphs. Internet Mathematics, 1(1): 1–35
9. Cafarella M, Cutting D (2004) Building Nutch: open source search. ACM Queue 2(2):54–61. doi: 10.1145/988392.988408
10. Choo KR, Smith RG (2008) Criminal exploitation of online systems by organised crime groups. Asian Journal of Criminology 3(1):37–59
11. Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. SIAM review 51:661–703. doi: 10.1137/070710111
12. Coleman C, Moynihan J (1996) Understanding crime data: haunted by the dark figure. Open University Press. ISBN 0-335-19519-9
13. Crucitti P, Latora V, Marchiori M, Rapisarda A (2003) Efficiency of scale-free networks: error and attack tolerance. Physica A: Statistical Mechanics and it Applications 320:622–642
14. Daley D, Kendall D (1964) Epidemics and rumours. Nature 204:1118. doi: 10.1038/2041118a0
15. Décary-Hétu D, Morselli C (2011) Gang presence in social network sites. International journal of cyber criminology 5(2):876–890
16. Efron B (2010) Large-scale inference: empirical Bayes methods for estimation, testing and prediction. Cambridge University Press
17. EUROPOL (2011) Internet facilitated organized crime. iOCTA (abridged). The Hague: European Police Office
18. Everitt B, Landau S, Leese M (2001) Cluster Analysis. Arnold, London
19. Ferri F, Grifoni P, Guzzo T (2012) New forms of social and professional digital relationships: the case of Facebook. Social network analysis and mining 2(2):121–137
20. Fisher R (1922) On the interpretation χ^2 from contingency tables, and the calculation of p. Journal of the Royal Statistical Society 85(1):87–94
21. Gallos LK, Barttfield P, Havlin S, Sigman M, Makse HA (2012) Collective behavior in the spatial spreading of obesity. Scientific Reports 2(45):1–9
22. Haynie D (2002) Friendship networks and delinquency: The relative nature of peer delinquency. Journal of Quantitative Criminology 18(2):99–134
23. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
24. Iribarren JB, Moro E (2009) Impact of human activity patterns on the dynamics of information diffusion. Physical Review Letters 103(3):8–11
25. Kantardzic M (2011) Data mining: concepts, models, methods, and algorithms. IEEE Press, Wiley-Interscience

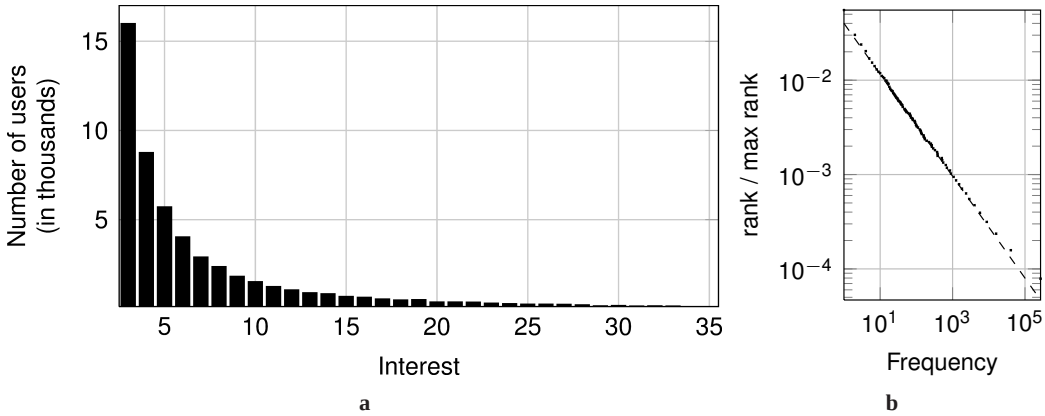


Fig. 6 a The histogram of interests expressed by the users in the crawled LiveJournal data set **b** The rank/frequency log-log plot of the histogram in 6a and the maximum likelihood power-law fit ($\gamma \approx 1.54$ and $x_{min} = 3$)

26. Lämmel R (2007) Google's MapReduce programming model — Revisted. Science of Computer Programming 70:1–30
27. Mityagin SA (2012) Modeling the spread of drug-addiction through the population on the basis of complex networks (in Russian - Modelirovanie processov narkotizatsiya nasileniya na osnove kompleksnix cetei). Dissertation, National Research University of Information Technologies, Mechanics and Optics (NRU ITMO)
28. Newman M (2005) Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46:323–351
29. Ochiai A (1957) A zoogeographic studies on the solenoid fishes found in Japan and its neighbouring regions. Bulletin of the Japanese Society of Fish Sciences 22:526–530
30. Onnela J, Saramäki J, Hyvönen J, Szabó G, Lazer D, Kaski K, Kertész J, Barabási AL (2007) Structure and tie strengths in mobile communication networks. Proceedings of the National Academy of Sciences of the United States of America 104(8):7332–7336
31. Pinto C, Mendes Lopez A, Machado J (2012) A review of power laws in real life phenomena. Communications in Nonlinear Science and Numerical Simulation 17(9):3558–3578
32. Porter MF (1980) An algorithm for suffix stripping. Program 14(3):130–137
33. Porter MF (2006) Stemming algorithms for various European languages. `http://www.snowball.tartarus.org/texts/stemmersoverview.html`. Accessed 19 November 2012.
34. Rhodes W, Kling R, Johnson P (2006) Using booking data to model drug user arrest rates: a preliminary to estimating the prevalence of chronic drug use. Journal of Quantitative Criminology 23(1):1–22
35. Scott J (2011) Social network analysis: developments, advances, and prospects. Social network analysis and mining 1(1):21–26
36. Suler J (2004) The online disinhibition effect. Journal of Cyberpsychology and Behaviour 7(3):321–326
37. Sunami AN (2007) Drug-conflict management in the context of information warfare (in Russian - Politika upravleniya narkokonfliktom v kontekste informatsionnoi voiny). Saint Petersburg State University.
38. Sutherland EH (1947) Principles of criminology. Philidelphia: JB Lippincott
39. Walsh C (2011) Drugs, the internet and change. Journal of Psychoactive Drugs 43(1):55–63
40. Williams P (2001) Organized crime and cybercrime: synergies, trends and responses. Arresting Transnational Crime 6(2)
41. Wilson RE, Gosling SD, Graham LT (2012) A review of Facebook research in the social sciences. Perspectives on Psychological Science 7(3):203–220
42. White T (2009) Hadoop: the definitive guide. O'Reilly Media, Yahoo! Press